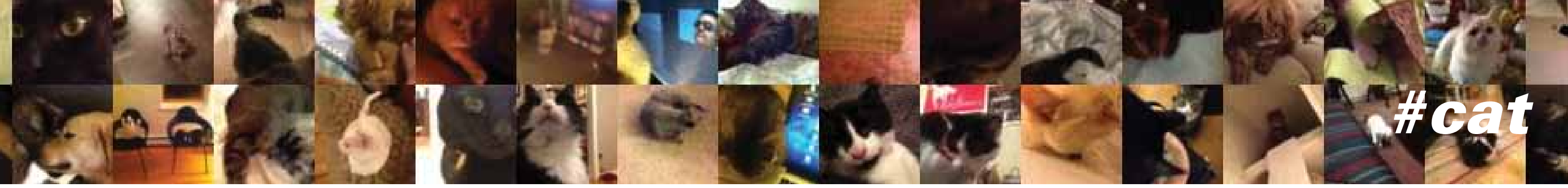
VISIONARY

how twitter is helping computers to see



HUMAN BEINGS ARE “VISIONARIES”: we rely on our vision, almost to the exclusion of our other senses, to inform us and guide our interactions with the world. In broad strokes, our vision gives us the ability to recognize *what*—objects and people in our environment, things that are similar, things that are different—as well as the ability to determine *where*—where things are, where they were, and where they will be.

Computers are not visionaries. Even when configured with superb optical “eyes” and abundant “brain” power, most computers see poorly, in the sense that, in non-controlled environments, their ability to determine *what* or *where* is limited and inconsistent. A computer would have a very difficult time identifying everything that would catch a human’s eye on a busy street. Furthermore, a computer optimized to recognize faces would likely fail miserably if asked to recognize a gun, or steer a car, or perform any visual task other than what it was optimized for.

Yet if current trends in computer technology, algorithm development, and data availability continue, computers will likely have excellent vision within the next five years. They will be able to process a complex visual scene quickly, accurately, and thoroughly, and will match or even outperform humans in most vision-specific tasks.

And helping to bring about that “see change” will be, of all things, Twitter, the social networking and microblogging service giant.

#Hey, look at this

It’s a little odd that humanity’s desire to blog is related to the pursuit of computer vision, but the connection exists because people often blog about what they see, and the computer has to be taught how to see. To recognize a cat, for example, the computer must first be given an image of a cat and told, “This is a cat.” In fact, the computer needs to have seen thousands of cats—in all positions, in different environments, at various angles, and under arbitrary light conditions such as a visual blog site might provide—so that it can recognize a cat regardless of circumstances.

Humans also have to learn to see, only the process is innate and happens largely without supervision. Learning begins essentially at birth and continues unabated for several years. By the end of its first year, an infant will have observed about a petabyte (10^{15} bytes) worth of data, or enough to fill 100,000 ten-gigabyte thumb drives. No computer has ever come close to being trained on so large a data set. Indeed, scientists’ inability to find a sufficiently large training set of annotated images has been a major stumbling block to realizing a sight-worthy computer.

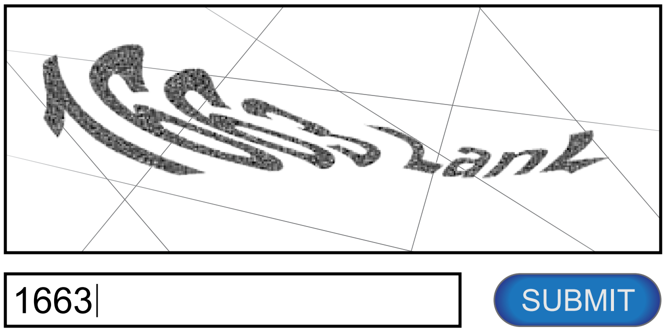
That changed on January 24, 2013, the day Twitter released Vine, an application that allows users to attach and send a short, six-second video tweet to followers. Vine has experienced exponential growth since then, and the amount of data sent collectively by Vine users has already topped 60 terabytes (60×10^{12} bytes). One couldn’t ask for a better training set.

“Millions of people are sending six-second videos to each other, each annotated with a statement like, ‘This is my cat playing,’ or ‘NYC’s best taco,’” says Steven Brumby of the Los Alamos National Laboratory, the lead scientist in an effort to develop a computer’s vision. “We can store the videos, then search the collection for “cat” and have ready access to hundreds of thousands of videos of cats. It’s a remarkable resource.”

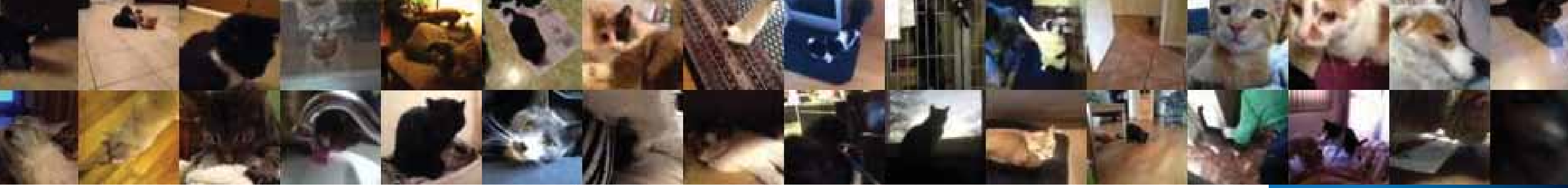
Couple the training set with a supercomputer able to execute several trillion operations per second (teraflops), and the goal of computer vision comes into view.

“Our focus is to develop the basic mathematics and computer science underpinning computer vision,” says Brumby. “I anticipate we’ll have visually adept computers within two years, in part because Google, Amazon, Silicon Valley startups, and several big academic groups are all working to make computer vision happen. This is the holy grail.”

It’s the holy grail because a sighted computer would enable a range of vital applications, foremost being autonomous robots that can be used for defense, manufacturing, resource extraction, emergency disaster response, environmental assessment, etc. If able to evaluate its environment faithfully, a seeing computer would usher in an era of computer-controlled transportation—non-stop trucking, coordinated traffic flow, and autonomous minivans that pick the kids up from soccer. Furthermore, a seeing computer would be an exceptional personal assistant, one with full access to the Internet and its body of knowledge that could help you keep tabs on your loved ones and watch over the sick and elderly.



Computers have such a hard time interpreting a visual field that a CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) has become a standard online security measure. Users are asked to extract information from a simple image and thus prove they are human beings.



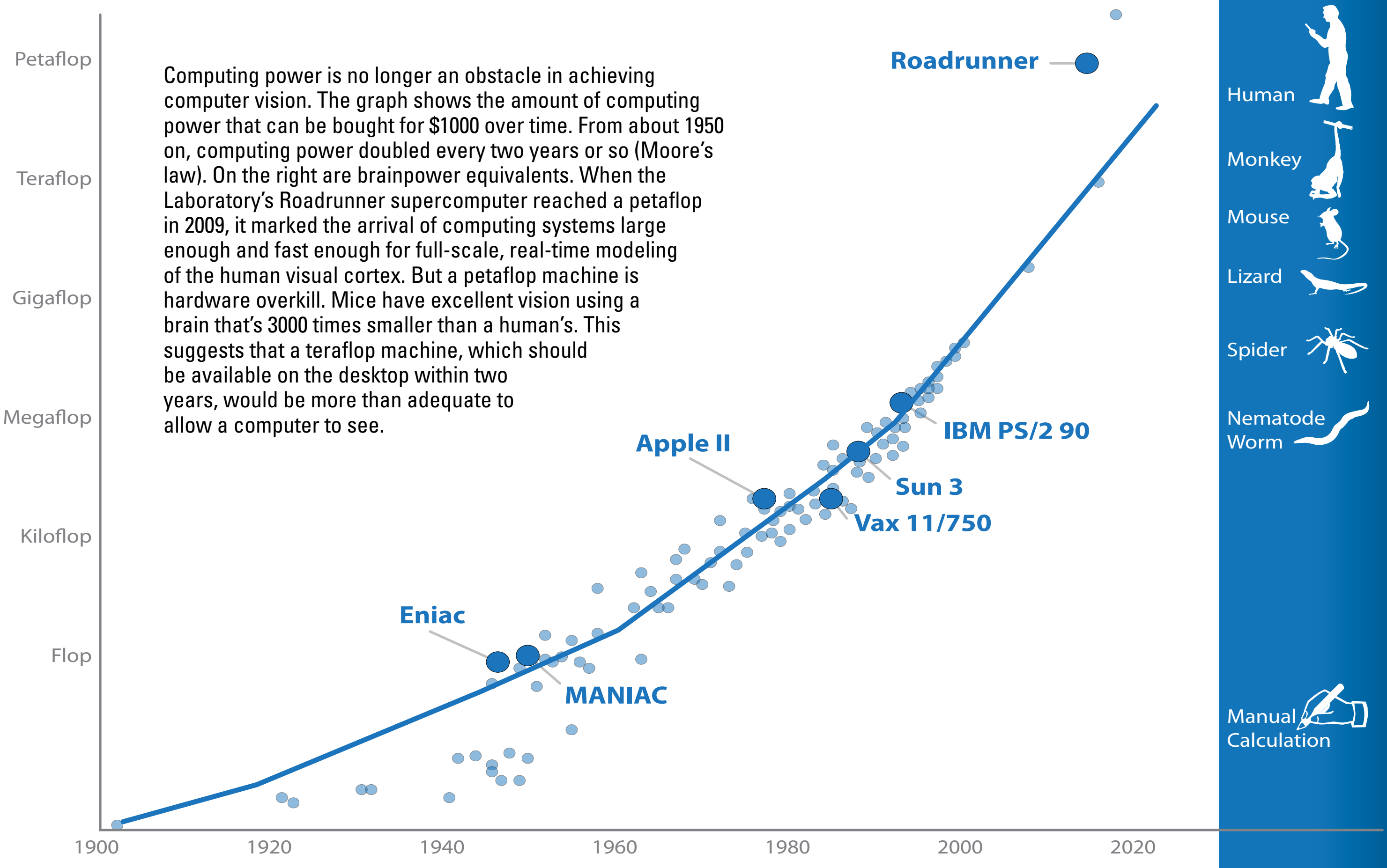
Unquestionably, a computer looking over your shoulder could be a good thing. But there are many who fear that a seeing computer will be the starting point of a sci-fi nightmare. Consider that the computer could use the camera in your smart phone or laptop, plus the network of traffic and security cameras that monitor essentially every street and alleyway of our cities, to identify you and the people around you and determine where you are and what you are doing. Private companies—for purely commercial reasons—are already beginning to master the technology for identifying and tracking consumers and their interests. Apart from raising issues of personal privacy, tracking can shift into surveillance, and the computer could be used to help achieve and sustain a police state.

But computers have been used and abused almost from the earliest days of computer applications. Today's powerful computers—the "predeprocessors" of which helped break Nazi communication codes during World War II—already monitor telecommunications and email in an effort to hunt down global terrorist groups. Despite this level of privacy intrusion, humanity has managed to thrive.

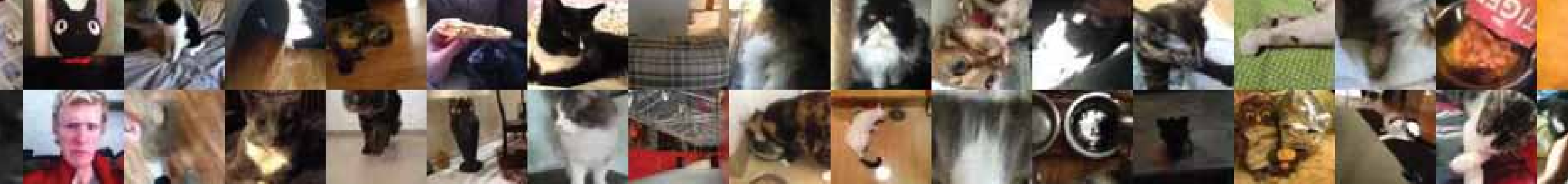
Let there be sight

How is a computer able to see? There is no simple answer, as currently the method pursued depends to a large extent on the visual task the computer will be performing, be it object recognition, event detection, video tracking, scene reconstruction, or something else. One area Los Alamos is pursuing is object recognition, basing its algorithms on models of how the human brain sees.

Briefly, object recognition starts by giving the computer a digital image that contains, say, a cat, and asking it to find all cats. The computer divvies the image, or portions of the image, into thousands, if not hundreds of thousands of tiny patches, with each patch being a tiny image perhaps 8 pixels by 8 pixels in size. The computer will try to represent, or duplicate, the information content of each patch by searching through a collection of patch-sized images that it has



Computing power is no longer an obstacle in achieving computer vision. The graph shows the amount of computing power that can be bought for $1000 over time. From about 1950 on, computing power doubled every two years or so (Moore's law). On the right are brainpower equivalents. When the Laboratory's Roadrunner supercomputer reached a petaflop in 2009, it marked the arrival of computing systems large enough and fast enough for full-scale, real-time modeling of the human visual cortex. But a petaflop machine is hardware overkill. Mice have excellent vision using a brain that's 3000 times smaller than a human's. This suggests that a teraflop machine, which should be available on the desktop within two years, would be more than adequate to allow a computer to see.



stored in memory, and selecting the ones that are a good match. The stored images are called features, and the collection of features is referred to as a dictionary.

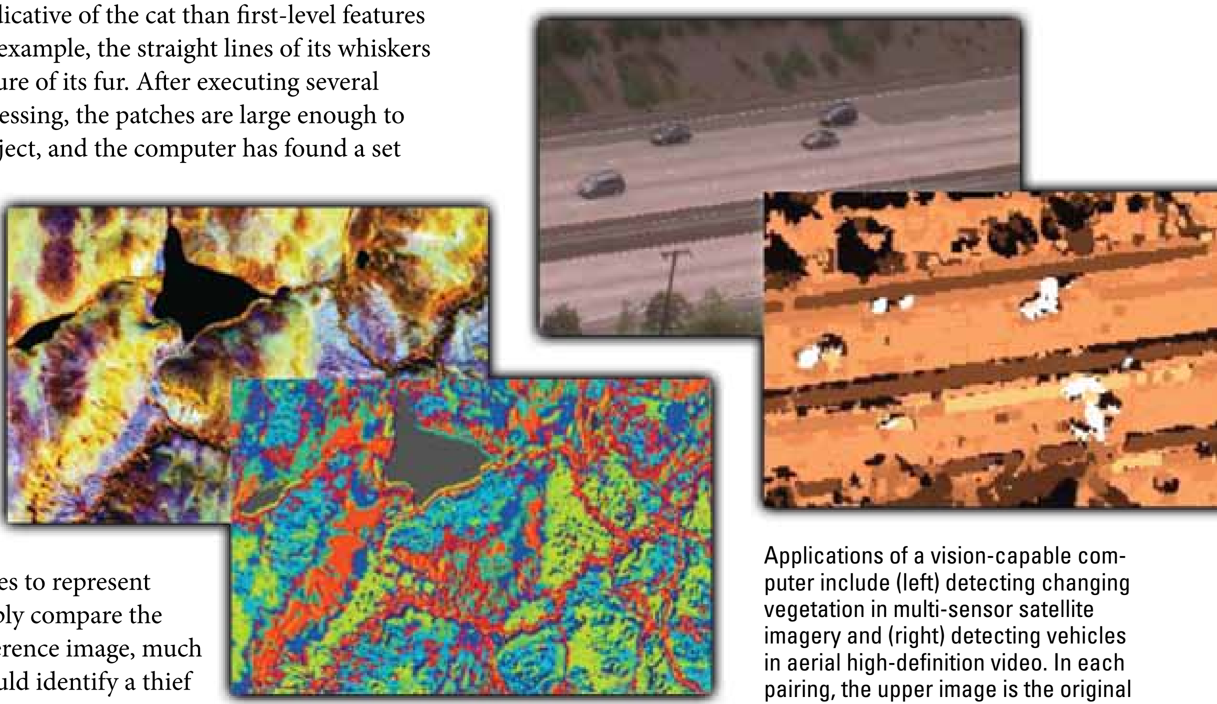
The features at this first level of processing are simple—a line pitched at a certain angle, a blotch of color, etc. But once the computer has done its best to represent each patch by combining one or more features, it moves on to the second processing level. The small patches are grouped together into larger patches that cover a greater fraction of the object. These larger patches are then represented by combinations of features contained in a second-level dictionary. These features are more indicative of the cat than first-level features and might show, for example, the straight lines of its whiskers or the color and texture of its fur. After executing several similar levels of processing, the patches are large enough to include the entire object, and the computer has found a set of features that accurately represent the object(s) in the input image. The set is given to a classification program, which plays a multi-dimensional game of Twenty Questions before it says, "It's a cat."

Why use features to represent objects, and not simply compare the entire image to a reference image, much the way a person would identify a thief by looking through a "dictionary" of mug shots? One reason is that the computer matches images by doing a pixel by pixel comparison and calculating a "distance parameter," with dissimilar images being farther apart.

Suppose there are two similar images of today's featured object, the cat, but in one, the cat's head is upright, while in the other, its tilted. A human would instantly recognize that its the same cat in both images, but the computer's pixel by pixel comparison would result in a large distance parameter, because a portion of the images don't line up. To obtain a closer match, the image dictionary would have to contain images of cats with their heads up, with their heads tilted, as well as every conceivable variation, so there would always have to be a stored image that aligns with the input image.

"Every object that we wanted the computer to recognize would need a similar portfolio of poses," says Brendt Wohlberg, a scientist working with Brumby. "The dictionary becomes extremely large and computationally very expensive to manipulate."

By breaking the image up into features, one can represent essentially any image by combinations of simpler images, much the way the entire English language can be constructed from combinations of just 26 letters.



Applications of a vision-capable computer include (left) detecting changing vegetation in multi-sensor satellite imagery and (right) detecting vehicles in aerial high-definition video. In each pairing, the upper image is the original observation, and the lower image is a computer-vision reconstruction.

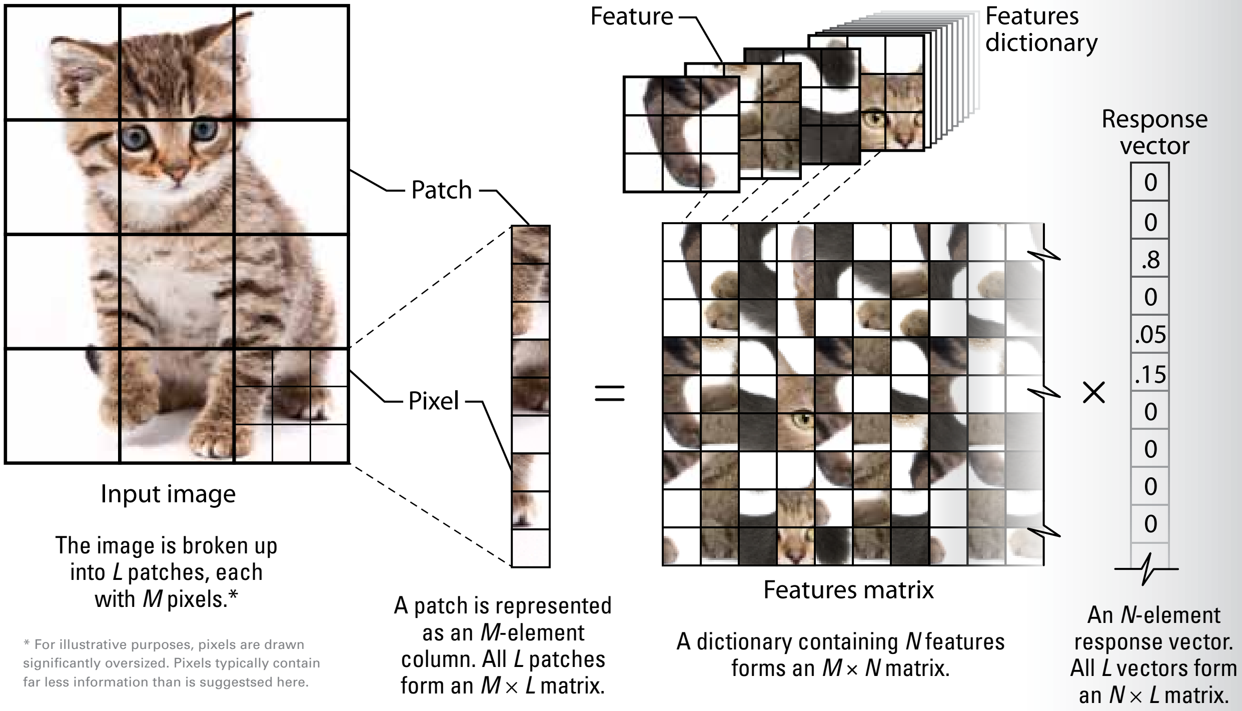
Processing an image, therefore, entails finding a way to represent hundreds of thousands of patches by combinations of just a few hundred features. Seeing in real time demands that the computer manipulate huge amounts of data—millions of pixels—very quickly. The processing rate needs to be teraflops or better. But the key to seeing lies in the features dictionaries, which need to have enough entries to represent the content of any patch. It's vital for the computer to have processed countless real-world images so that it can build its dictionary.

Sparse Training

“Natural images, or portions of a natural image, are generally not that complex,” says Rick Chartrand, who shares lead investigator duties with Steven Brumby. “They can almost always be represented by a sparse representation of features, assuming one has a sufficient number of features to choose from.”

Sparse representations, which figure heavily in the computer-vision algorithms being developed at Los Alamos, can be understood with a little help from a mixed drink. Consider a local bar, stocked with dozens of liquors, juices, sodas, waters and flavorful mixes—the full inventory forming a “liquids dictionary.” Any drink can be made by mixing the elements in the dictionary together in various amounts. For example, a gin and tonic is made with 1 part gin, 3 parts tonic water, and 0 parts of every other liquid in the bar. Each drink is a sparse representation of the liquids dictionary, in that each is made by mixing together only a few of the dictionary’s many elements.

Matrix equation for a single patch



Input image

The image is broken up into L patches, each with M pixels.*

A patch is represented as an M-element column. All L patches form an $M \times L$ matrix.

Features matrix

A dictionary containing N features forms an $M \times N$ matrix.

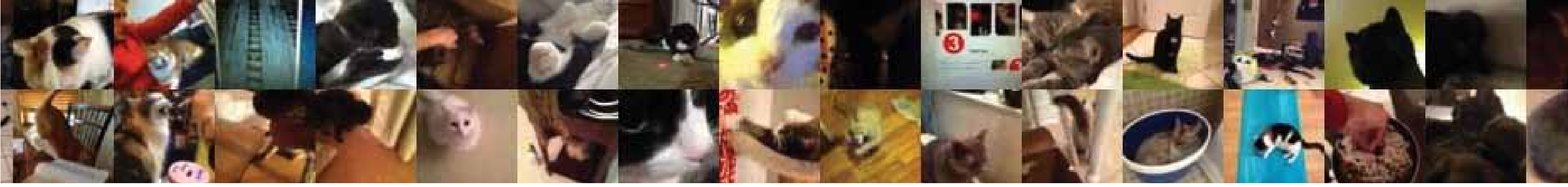
An N-element response vector. All L vectors form an $N \times L$ matrix.

* For illustrative purposes, pixels are drawn significantly oversized. Pixels typically contain far less information than is suggestsed here.

The computer recognizes an object by finding a sparse representation of it. The image is broken up into perhaps a hundred thousand tiny images called patches. Features are patch-sized color images that the computer can recall from a “dictionary” stored in memory. Like mixing a drink, just about any patch can be reproduced by mixing a few features together in various amounts. The amounts are encoded in an entity called a response vector.

The goal is to find the mix of features that accurately reproduces a patch, which entails finding the optimal reponse vector. Each patch or feature is made up of pixels, and by rearranging the pixels, one can represent the patches, the features dictionary, and the response vectors as matrices in a matrix equation. The optimal solution to the equation will be a set of sparse response vectors, each one a recipe for reproducing a patch as a sparse representation of the features dictionary.

In training a computer to see, a large set of, say, 500 images is processed simultaneously. At 100,000 patches per image, all patches form a matrix of 50 million columns, and the computer uses the same features dictionary to find the 50-million-column matrix of response vectors. Any of those solutions can be added to the dictionary to improve it. The computer can thus bootstrap and optimize both its dictionary and response vectors. The more images a computer processes, the better its features dictionaries and the better its sight.



On the corner of Twitter and Vine

The six-second Vine videos that can be attached to Twitter tweets are a computer-vision resource the likes of which the world has never encountered. The phone app is easy to use, and people have responded, filming and posting to the world microvideos of their cat at play, the friendly waiter at the restaurant, local street performers, or baby's first steps. When terrorist bombs exploded during the 2013 Boston Marathon, local Vine users posted video of the chaos almost instantaneously, while users everywhere took, then forwarded, video of television newscasts, spreading word of the disaster at an unprecedented pace. Access to the videos is free—they are public domain and any Twitter user can download them—making what happened in Boston available for public scrutiny and analysis.

When Los Alamos cosmologist and computer sensei Mike Warren, who also works with Brumby and Wohlberg, heard about the Vine release, he immediately recognized the potential to create a unique resource for vision research. Utilizing a prototype storage system (initially developed to archive astronomy data and the results of supercomputer simulations), he wrote software to download and archive the videos. The stream of data he started collecting in the early spring has since become a deluge. Warren estimates that during peak Vine usage this summer he was collecting more than a million videos per day.

A quick perusal of the data reveals an unrivaled training set. For example, searching the tweets for videos annotated with the word "cat" finds more than 250,000 videos, most with at least one cat in it. (A similar search finds more than 400,000 videos of dogs, apparently the Vine user's BFF). Selecting videos based on their dominant color, like green or blue, reveals thousands of short films showing grass or sky. Stills from the videos can be used for training, or the video themselves can be used to teach the computer to detect motion.

"Watching 24 hours a day, it would take 12 years to view the video we have now. That's an amount of information that rivals everything an 18-year-old has ever seen or heard," said Warren.

Where it stands

Life at a national nuclear security laboratory is a little different, in that security is a priority and, one way or another, affects every process and procedure. The institutional supercomputers that will be taught to see were ill-equipped to receive an ocean of unknown, unverified data downloaded from the Web, and system engineers and cybersecurity experts have yet to resolve the myriad of throughput and security issues. Only a tiny fraction of Vine data has been processed, and a remarkably patient Warren waits for whatever changes that need to be made to be made.

Undeterred, Brumby's team is continuing to explore different algorithms that will process more data faster and achieve a higher level of recognition fidelity. They are also refining methods that enable a computer to search through an enormous data set unsupervised, so that it learns on its own which features to extract to best help it identify objects.

Steven Brumby with a Vine backdrop filtered by the keyword "blue."

Will a computer ever truly be able to see? If sight is simply extracting information from light, then computers are already seeing, and Brumby and his team can be viewed as high-powered ophthalmologists working to make computers see better. But seeing is often tied to awareness, to interpreting the light-based information so as to understand the world around us. While cognizant computers are a staple of science fiction, they are at present not part of the real world. It's doubtful a computer will ever see things the way we do. But whether human-like or not, computers will attain excellent vision within our lifetime, and the world around us will be forever changed. LDRD

—Jay Schecker